Assessing Gene-Environment Interactions for Common and Rare Variants with Binary Traits Using Gene-Trait Similarity Regression

Guolin Zhao,* Rachel Marceau,* Daowen Zhang,* and Jung-Ying Tzeng*,†,[1]
*Department of Statistics and †Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695

ABSTRACT Accounting for gene–environment ($G\times E$) interactions in complex trait association studies can facilitate our understanding of genetic heterogeneity under different environmental exposures, improve the ability to discover susceptible genes that exhibit little marginal effect, provide insight into the biological mechanisms of complex diseases, help to identify high-risk subgroups in the population, and uncover hidden heritability. However, significant $G\times E$ interactions can be difficult to find. The sample sizes required for sufficient power to detect association are much larger than those needed for genetic main effects, and interactions are sensitive to misspecification of the main-effects model. These issues are exacerbated when working with binary phenotypes and rare variants, which bear less information on association. In this work, we present a similarity-based regression method for evaluating $G\times E$ interactions for rare variants with binary traits. The proposed model aggregates the genetic and $G\times E$ information across markers, using genetic similarity, thus increasing the ability to detect $G\times E$ signals. The model has a random effects interpretation, which leads to robustness against main-effect misspecifications when evaluating $G\times E$ interactions. We construct score tests to examine $G\times E$ interactions and a computationally efficient EM algorithm to estimate the nuisance variance components. Using simulations and data applications, we show that the proposed method is a flexible and powerful tool to study the $G\times E$ effect in common or rare variant studies with binary traits.

KEYWORDS binary traits; gene–environment interaction; rare variant association; GLMM; marker-set interaction analysis; variance-component methods

HUMAN complex traits have a multifactor etiology that involves the interplay between genetic susceptibility and environmental exposures. Studies of gene–environment ($G\times E$) interactions can facilitate our understanding of genetic heterogeneity under different environmental exposures (Kraft *et al.* 2007; Van Os and Rutten 2009), help to identify high-risk subgroups in the population (Murcray *et al.* 2009), provide insight into the biological mechanisms of complex diseases (Thomas 2010), and improve the ability to discover susceptible genes that interact with other factors but exhibit little marginal effect (Thomas 2010). However, finding significant $G\times E$ interactions is not an easy task. Model misspecification, inconsistent definitions of the environmental variable, and insufficient sample sizes are just a few of the issues that often lead to low power and nonreproducible findings in $G\times E$ studies (Mechanic *et al.* 2012; Jiao *et al.* 2013; Winham and Biernacka 2013). In particular, the sample size needed to detect a $G\times E$ effect is usually four times larger than that needed to detect a main effect of similar magnitude (Thomas 2011). Thus, researchers need a robust, powerful $G\times E$ test to generate reproducible findings.

Conventionally, researchers search for significant genetic or $G\times E$ associations, using single-SNP methods, *e.g.*, the Kraft 2-d.f. test (Kraft *et al.* 2007) or the simultaneous test of Dai (Dai *et al.* 2012). More complex methods (*e.g.*, Mukherjee and Chatterjee 2008; Murcray *et al.* 2009; Sohns *et al.* 2013) aim to retain the advantages from both the case-only test (high power but sensitive to G–E correlations) and the standard case–control $G\times E$ test (low power but robust to G–E correlations). Despite the many efforts to improve single-SNP $G\times E$ tests, issues remain; *e.g.*, a large proportion of trait heritability remains unexplained (Manolio *et al.* 2009) due to false positive and/or false negative findings.

Inflated false positive rates arise when the model used to screen for $G\times E$ interactions does not correctly reflect the true underlying genetic (G) and environmental (E) effects (Voorman *et al.* 2011; Lin *et al.* 2013; Wang *et al.* 2013). To address this issue, Voorman *et al.* (2011) suggested a model-robust estimate of the variance, and Lin *et al.* (2013) and Wang *et al.* (2013) suggested a random-effect model to capture

doi: 10.1534/genetics.114.171686
Manuscript received October 12, 2014; accepted for publication December 29, 2014; published Early Online January 12, 2015.
[1] Corresponding author: Department of Statistics, North Carolina State University, Campus Box 7566, Raleigh, NC 27695. E-mail: jytzeng@stat.ncsu.edu

the genetic main effect. The false negative (underpower) issues can be addressed by evaluating $G{\times}E$ effects on a set of markers, *e.g.*, on genes, linkage disequilibrium (LD) blocks, or pathways (Tzeng *et al.* 2011; Lin *et al.* 2013). Marker–set $G{\times}E$ analysis can improve power by aggregating effects across markers. Such accumulation methods account for LD among markers and reduce the total number of tests to be performed. The improved power is particularly crucial for common variants with subtle individual effects and for rare variants with sparse occurrence (Sham and Cherny 2011). In addition, operating at a gene/pathway level helps increase reproducibility (Sohns *et al.* 2013).

Several $G{\times}E$ marker-set methods are available to study associations with common variants, where the major task is to avoid a large number of parameters for modeling G, E, and $G{\times}E$ variables. One of the first proposed $G{\times}E$ marker-set methods was Tukey's 1-d.f. test (Chatterjee *et al.* 2006), which made significant progress toward fully understanding complex diseases. However, this method makes the often incorrect assumption that a SNP's interaction effect is proportional to its marginal genetic effect (Winham and Biernacka 2013). Other commonly adopted $G{\times}E$ marker-set methods include minimum P-value (min-P) methods and weighted burden methods, where weights can be obtained from the principal components (PCs) of the SNP genotypes (Winham and Biernacka 2013) or from the G–E correlation (Jiao *et al.* 2013). In particular, Jiao *et al.* (2013) showed that the correlation between G and E can serve as an informative indicator for $G{\times}E$ interactions and that incorporating G–E correlations as weights can increase the signal-to-noise ratio in a $G{\times}E$ marker set while avoiding permutations. However, these observations are valid only when the true G–E correlation is in the same direction as the $G{\times}E$ interaction (Jiao *et al.* 2013). Fan and Lo (2013) proposed a model-free approach based on a summation of partitions to evaluate the interaction effects for rare variants. However, their method evaluates only the combined effect of G and $G{\times}E$, not the separated effects. Recently, Lin *et al.* (2013) proposed a generalized linear mixed-effect model (GLMM) for $G{\times}E$ interactions for binary and continuous traits and showed it has superior power and robustness over min-P methods. A similar method, similarity regression (SimReg), proposed by Tzeng *et al.* (2011) to study marker-set $G{\times}E$ for continuous traits, was shown to be connected to linear mixed-effect models.

In this article, we extend the SimReg $G{\times}E$ framework established in Tzeng *et al.* (2011) to binary traits with common or rare variants. SimReg, which is inspired by Haseman–Elston regression for linkage analysis (Haseman and Elston 1972; Elston *et al.* 2000) and haplotype similarity tests for regional association (Tzeng *et al.* 2003; Beckmann *et al.* 2005), uses a regression model to correlate trait similarity with genetic similarity across multiple loci and to account for covariates. SimReg has been shown to perform well for common and rare variants (Tzeng *et al.* 2011). However, unlike similarity-based testing for the genetic main effect (Tzeng *et al.* 2009) or for $G{\times}E$ with quantitative traits (Tzeng *et al.* 2011), $G{\times}E$ tests with binary traits have several challenges associated with computation and estimation. In particular, $G{\times}E$ tests require the estimation of nuisance parameters to capture the main effects. Estimating these parameters requires high-dimensional integration and the inversion of a high-dimensional similarity matrix. For quantitative $G{\times}E$ tests, this estimation can be sidestepped using the normality of the phenotype, but no such useful properties exist for binary $G{\times}E$ tests. To overcome these challenges, Lin *et al.* (2013) proposed using ridge regression to estimate the nuisance main effects, selecting the tuning parameter using generalized cross validation.

In our work, we develop an EM algorithm to approximate the integration and we alleviate the computational burden of maximum-likelihood estimation (MLE) by performing a low-rank approximation of the similarity matrix. We show that the SimReg coefficient can be expressed as a variance component of a working GLMM, which facilitates the derivation of a test statistic and unifies SimReg with other random-effect-based methods (*e.g.*, Lin *et al.* 2013; Wang *et al.* 2013). The proposed SimReg method can incorporate covariates and uses a permutation-free procedure to evaluate $G{\times}E$ effects. In addition, the proposed method extends the model from linear effects (*e.g.*, Jiao *et al.* 2013; Lin *et al.* 2013) to other complex effects by selecting appropriate similarity metrics, and it avoids the need to select tuning parameters. Unlike current robust marker-set $G{\times}E$ methods that focus on common variant analysis, we investigate the performance of the proposed $G{\times}E$ strategy with rare and common variants. We evaluate the validity and power of the proposed method using simulation studies and illustrate the utility of the proposed method via two data applications: one studies the interactions between *PLA2G7* and physical activity on obesity, using Cohorte Lausannoise (CoLaus) sequencing data, and a second assesses the effect modifier role of body mass index (BMI) on the association between *TCF7L2* and type 2 diabetes, using the Wellcome Trust Case Control Consortium data.

Materials and Methods

Gene–trait similarity regression for G×E effects

Let Y_i be the binary disease indicator for individual i ($i = 1, \ldots, n$); *i.e.*, $Y_i = 1$ if individual i has the disease of interest and $Y_i = 0$ otherwise. Let G_i^m be the minor allele count for individual i at locus m ($m = 1, \ldots, M$), let X_{Ei} be a $1 \times K_E$ vector of environmental factors, and let X_{Ci} be the $1 \times K_C$ vector of confounders. The full covariate vector is $X_i = (1, X_{Ci}, X_{Ei})$ with dimension $1 \times (1 + K_C + K_E)$. All covariates are standardized to have a mean of 0 and a variance of 1. For illustration, we consider the case where $K_E = 1$, but it is straightforward to extend the proposed work to $K_E > 1$.

We quantify the trait similarity for a pair of individuals i and j, T_{ij}, as the weighted sample covariance between their disease statuses; *i.e.*, $T_{ij} = \{\omega_i(Y_i - \mu_i^0)\}\{\omega_j(Y_j - \mu_j^0)\}$, where $\mu_i^0 = E(Y_i|X_i)$ is the subject-specific trait mean accounting for covariate X_i but assuming no genetic effects and ω_i is a weight accounting for the fact that the Y_i's have difference variances (Tzeng *et al.* 2009). From this definition, the expected trait similarity $E(T_{ij}X) = \omega_i\omega_j \times E\{(Y_i - \mu_i^0)(Y_j - \mu_j^0)\}$ is the covariance of Y_i and Y_j with weights $\omega_i\omega_j$. For binary traits, we assume a logistic model, $\mu_i^0 = e^{X_i\gamma}/(1 + e^{X_i\gamma})$, where γ is the coefficient vector of the covariate X_i and $\omega_i = \mu_i^0(1 - \mu_i^0)$ is the optimal weight for the logistic model (Tzeng *et al.* 2009).

Genetic similarity is calculated as the weighted sum of single-marker similarities; *i.e.*, $S_{ij} = \sum_{m=1}^{M} w_m s(G_i^m, G_j^m)$, where $s(G_i^m, G_j^m)$ is the genetic similarity at marker m and w_m is the weight. There are several choices for $s(G_i^m, G_j^m)$ (*e.g.*, Wessel and Schork 2006; Schaid 2010a); a popular one is the identity-by-state (IBS) metric: $s_{\mathrm{IBS}} = 2 - \left|G_i^m - G_j^m\right|$. Weights w_m are typically based on allele frequencies, the degree of evolutionary conservation, or the functionality of the variants (Wessel and Schork 2006; Price *et al.* 2010; Schaid 2010a,b). For example, one can use the minor allele frequency (MAF) of marker m, denoted by q_m, to up-weight similarities that are contributed by rare variants: *e.g.*, $w_m = (1 - q_m)^{24}$ (Wu *et al.* 2011) can be used to target rare variants only, or a moderate weight $w_m = q_m^{-3/4}$ (Pongpanich *et al.* 2012) can be used to promote similarities attributed to rare alleles while retaining the contributions from common variants.

The proposed $G{\times}E$ gene–trait similarity regression model is

$$E(T_{ij}|X,S) = a + b \times X_{Ei}X_{Ej} + c \times S_{ij} + d \times S_{ij} \times X_{Ei}X_{Ej}, \quad i \neq j. \tag{1}$$

Because T_{ij} incorporates baseline covariate information, model (1) does not contain an intercept or an $X_{Ei}X_{Ej}$ interaction covariate term (*i.e.*, $a = b = 0$) (Tzeng *et al.* 2011). Using model (1), one can assess the $G{\times}E$ interaction by testing $H_0^{GE}: d = 0$, or one can perform a joint test for the genetic main effect and $G{\times}E$ interactions simultaneously by testing $H_0^{\text{Joint}}: c = d = 0$. The joint test is recommended if either the genetic heterogeneity or the $G{\times}E$ interaction mechanism is unknown (Kraft *et al.* 2007; Tzeng *et al.* 2011).

Score test for G×E effects and joint effects

Following a similar procedure to that found in Tzeng *et al.* (2009), we connect the similarity regression to a working GLMM to derive the score test. Consider the following GLMM,

$$g(\mu) = X\gamma + h_G + h_{GE}, \tag{2}$$

where $\mu = (\mu_1, \ldots, \mu_n)$ is a vector of conditional means $\mu_i = E(Y_i|X, h_G, h_{GE})$ and $g(.)$ is a link function. Here, we consider a logit link $\mathrm{g}(\mu_i) = \log\ \{\mu_i/(1 - \mu_i)\}$. Vectors $h_{G(n\times 1)} = (h_{G1}, \ldots, h_{Gn})$ and $h_{GE(n\times 1)} = (h_{GE1}, \ldots, h_{GEn})$ contain the subject-specific genetic main effect and $G{\times}E$ interaction, respectively. Assume h_G and h_{GE} are random effects; *i.e.*, $h_G \sim N(0, \tau_G S_G)$ and $h_{GE} \sim N(0, \tau_{GE} S_{GE})$ with $S_G = \{S_{ij}\}$, $S_{GE} = DS_GD$, and $D = \mathrm{diag}\{X_{Ei}\}$. Then, the marginal covariance of Y_i and Y_j in this working model is

$$\mathrm{cov}(Y_i, Y_j) \approx \left\{g'\left(\mu_i^0\right)g'\left(\mu_j^0\right)\right\}^{-1} \times \{\tau_G S_{ij} + \tau_{GE} X_{Ei}X_{Ej}S_{ij}\},$$

where $g'(\mu) = \partial g(\mu)/\partial\mu$ (see *Appendix A*). Recall the expected trait similarity is $E(T_{ij}|X) = \omega_i\omega_j \times \mathrm{cov}(Y_i, Y_j)$. Therefore,

$$\begin{aligned} E(T_{ij}|X) &\approx \omega_i\omega_j \times \left\{g'\left(\mu_i^0\right)g'\left(\mu_j^0\right)\right\}^{-1} \\ &\quad \times \{\tau_G \times S_{ij} + \tau_{\mathrm{GE}} \times X_{Ei}X_{Ej}S_{ij}\} \\ &= \tau_G \times S_{ij} + \tau_{\mathrm{GE}} \times X_{Ei}X_{Ej}S_{ij}, \end{aligned}$$

where $\omega_i = g'(\mu_i^0) = 1/\mu_i^0(1 - \mu_i^0)$. In other words, we can examine $H_0^{GE}: d = 0$ and $H_0^{\text{Joint}}: c = d = 0$ of model (1) by testing $H_0^{GE}: \tau_{GE} = 0$ and $H_0^{\text{Joint}}: \tau_G = \tau_{GE} = 0$ in model (2), respectively.

To derive the score test statistics, we rewrite model (2) as

$$g(\mu) = X\gamma + Z_G b + Z_{GE} b_{GE}, \tag{3}$$

where $b \sim N(0, \tau_G I_{L\times L})$, $b_{GE} \sim N(0, \tau_{GE} I_{L\times L})$, L is the rank of matrix S_G, and Z_G is a $n \times L$ matrix satisfying $Z_G Z_G^T = S_G$. Matrix Z_{GE} is defined in the same manner as Z_G, and $Z_{GE} = DZ_G$ because $S_{GE} = DS_GD$. Following Zhang and Lin (2003), the score statistic to examine the $G{\times}E$ effect (*i.e.*, testing $H_0^{GE}: \tau_{GE} = 0$) can be calculated as

$$\begin{aligned} &U_{GE} \\ &= \frac{1}{2}\left\{(y_1^W - X\gamma)^T V_1^{-1} S_{GE} V_1^{-1} (y_1^W - X\gamma) - tr(P_1 S_{GE})\right\}\Big|_{\tau_G = \widehat{\tau_G}, \tau_{GE} = 0, \gamma = \hat{\gamma}}, \end{aligned}$$

where $y_1^W = X\hat{\gamma} + Z_G\hat{b} + \Delta_G(y - \hat{\mu}^G)$ is the working vector in model (3) under $H_0^{GE}: \tau_{GE} = 0$; $\mu^G = E(YX, b) = g^{-1}(X\gamma + Z_G b)$; $\Delta_G = \mathrm{diag}\{g'(\mu_i^{\mathrm{G}})\}$ with $g'(\mu_i) = 1/\{\mu_i(1 - \mu_i)\}$, and μ_i^G is the ith entry of μ^G; $\widehat{\tau_G}$ and $\hat{\gamma}$ are the MLEs for τ_G and γ under H_0^{GE}, respectively; $V_1 = W_G^{-1} + \tau_G S_G$ with $W_G = \mathrm{diag}\{\mu_i^G(1 - \mu_i^G)\}$, and $P_1 = V_1^{-1} - V_1^{-1}X(X^T V_1^{-1} X)^{-1} X^T V_1^{-1}$. As noted in the literature (Zhang and Lin 2003; Tzeng and Zhang 2007), the second term, $tr(P_1 S_{GE})$, is the mean of the first term and its variability is small compared to the first term. Thus, we derive our test statistic using only the first term; *i.e.*,

$$T_{GE} = \frac{1}{2}\left\{(y_1^W - X\gamma)^T V_1^{-1} S_{GE} V_1^{-1} (y_1^W - X\gamma)\right\}\Big|_{\tau_G = \widehat{\tau_G},\ \tau_{GE} = 0,\ \gamma = \hat{\gamma}}.$$

We propose an EM algorithm in *Appendix B* to obtain the MLEs for τ_G and γ.

In a similar manner, the score statistic under $H_0^{\text{Joint}}: \tau_G = \tau_{GE} = 0$ can be obtained as

$$\begin{aligned} U_{\text{Joint}} = \frac{1}{2}\Big\{&(y_0^W - X\gamma)^T V_0^{-1}(S_{GE} + S_G)V_0^{-1}(y_0^W - X\gamma) \\ &- tr[P_0(S_{GE} + \mathrm{S_G})]\Big\}\Big|_{\tau_G = 0,\ \tau_{GE} = 0,\ \gamma = \tilde{\gamma}}, \end{aligned}$$

and we define the test statistic of the joint effect as

$$T_{\text{Joint}} = \frac{1}{2}\Big\{(y_0^W - X\gamma)^T V_0^{-1} \times (S_{GE} + S_G) V_0^{-1} (y_0^W - X\gamma)\Big\}\Big|_{\tau_G=0,\ \ \tau_{GE}=0,\ \ \gamma=\tilde{\gamma}},$$

where $y_0^W = X\tilde{\gamma} + \Delta\left(y - \widehat{\mu^0}\right)$ is the working vector under $H_0^{\text{joint}}: \tau_G = \tau_{GE} = 0$. Here, $\mu^0 = E(YX) = g^{-1}(X\gamma)$, $V_0 = W_0^{-1}$, $W_0 = \text{diag}\{\mu_i^0(1-\mu_i^0)\}$, $P_0 = V_0^{-1} - V_0^{-1}X(X^TV_0^{-1}X)^{-1}X^TV_0^{-1}$, and $\tilde{\gamma}$ is the MLE for γ under H_0^{Joint}.

We show in *Appendix C* that T_{GE} and T_{Joint} follow a weighted χ^2-distribution asymptotically under H_0^{GE} and H_0^{Joint}, respectively. *P*-values can then be calculated numerically using moment-matching approximations (Duchesne and Lafaye de Micheaux 2010).

Low-rank approximation of S_G for computational and statistical efficiency

The calculation of the G×E test statistic involves the inversion of matrices V_1 and S_{GE}, both of dimension $n \times n$. When n is large (*e.g.*, >5*k*), direct inversion of these matrices can be computationally intensive, and the inversion must be performed at every EM iteration to obtain main-effect term b (see *Appendix B*). To reduce the computational intensity and to facilitate the inversion of these matrices, we consider a low-rank approximation of S_G. The low-rank approximation has been used in the literature to improve power when the number of markers increases and when more noise is incorporated into S_G (Cai *et al.* 2011). Previous works (Cai *et al.* 2011; Tzeng and Zhang 2007; Tzeng *et al.* 2011) indicate that S_G is a positive semidefinite matrix, for which there are a few dominant eigenvalues. Assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\tilde{L}}$, $\tilde{L} \leq L$, are the leading eigenvalues that explain the majority of the variance of S_G [*i.e.*, $\Sigma^{\hat{L}}_{\ell=1}\lambda_\ell / \Sigma^L_{\ell=1}\lambda_\ell \geq p$ for some $p \in (0,1]$] and have corresponding eigenvectors $e_1, e_2, \cdots, e_{\tilde{L}}$. Then, we approximate Z_G by $\widetilde{Z_G} \equiv \left[\sqrt{\lambda_1}e_1, \ldots, \sqrt{\lambda_{\tilde{L}}}e_{\tilde{L}}\right]$ For an appropriate choice of p (*e.g.*, $p = 0.90 \sim 0.99$), $\widetilde{S_G} = \widetilde{Z_G}\widetilde{Z_G}^T$ contains most of the information from S_G. Especially with rare variant data, $\tilde{L}$ is usually $< L$, and the computation is more straightforward.

Miao (2009) indicated that the potential bias caused by a low-rank approximation can be minimized if a high percentage of the variation of S_G can be retained. In our explorations, we found that selecting too small a p did not affect the test size but did lead to power loss because too much genetic information is discarded. We also found that the power loss with a large p (*e.g.*, $p = 0.99$) was negligible but could stabilize the numerical calculation and boost computational efficiency. The improvement when $p = 0.99$ occurs because S_G has many eigenvalues that are near zero. Using a p slightly <1 removes a large number of near-zero eigenvalues, which stabilizes the numerical computations, shortens the computational time, and yields a type I error rate close to the nominal level (Table 1).

Simulation studies

To investigate the performance of the proposed SimReg G×E method, we conducted simulation studies. The first simulation focuses on rare-variant (RV) analysis using sequence data, and the second simulation focuses on common-variant (CV) analysis using HapMap data. The simulation data and code are available from the Dryad Digital Repository (http://datadryad.org/) at http://doi.org/10.5061/dryad.742gv (*i.e.*, Dryad data identifier:doi:10.5061/dryad.742gv).

RV simulations: We obtained 10,000 haplotypes for a 1-Mb region simulated by COSI (Schaffner *et al.* 2005) according to a coalescent model where the LD pattern and population history mimicked those of the European population. We selected the first 100 rare loci [*i.e.*, minor allele frequency (MAF) <5%] for further analyses. We randomly drew 2 haplotypes with replacement from the 10,000 to form each subject's genotype. We generated the binary phenotype from a Bernoulli (π_i) distribution, where $\pi_i = e^{\eta_i}/(1+e^{\eta_i})$, $\eta_i = \gamma_0 + X_{Ei}\gamma_E + \sum_{r=1}^R G_{ri}\gamma_G^r + \sum_{r=1}^R G_{ri}X_{Ei}\gamma_{GE}^r$, R is the number of causal loci, and G_{ri} is the number of rare alleles at causal locus r, $1 \leq r \leq R$. While we varied the value of R, we controlled the population attributable risk (PAR) at a_G and a_{GE} for the genetic main effect and G×E effect, respectively (Madsen and Browning 2009). Given a_G, a_{GE}, and R, we calculate γ_G^r and γ_{GE}^r using $\gamma_G^r = \log\{(a_G/R)/((1-a_G/R) \times q_r) + 1\}$ and $\gamma_{GxE}^r = \log\{(a_{GE}/R)/\ \ ((1-\alpha_{GE}/R) \times q_r) + 1\}$ (Madsen and Browning 2009), where $r = 1, \ldots, R$, and q_r is the MAF for the *r*th locus based on the 10,000 haplotypes. We considered both case–control sampling with 750 cases and 750 controls and random sampling with sample size 1500 and prevalence rate 0.3.

In the type I error analysis, we set $(a_G, a_{GE}) = (0,0)$ for the joint test and considered $(a_G, a_{GE}) = (0,0)$ and $(0.02, 0)$ for the G×E test. Because the burden-based tests are sensitive to the misspecification of the main-effect model (Voorman *et al.* 2011), we set a weak main-effect PAR so that the burden-based tests can still serve as a valid benchmark. We performed 10,000 replicates per scenario. In the power analysis, we set $(a_G, a_{GE}) = (0.02, 0.1)$ for both the G×E test and the joint test and considered $R = 20,\ 40,\ 60,\ 80,$ and 100. We performed 500 replicates per scenario. In all analyses, the 100 loci were included in the association tests.

SimReg's performance was compared to GESAT (Lin *et al.* 2013) and a burden-based G×E test. GESAT is a GLMM-based G×E test that is closely connected to SimReg: from the GLMM representation in model (2), we see that SimReg assumes $h_{GE} \sim N(0,\ \tau_{GE}S_{GE})$, where S_{GE} (calculated through the similarity kernel) determines how the G×E effects are modeled. In contrast, GESAT assumes a linear effect on h_{GE}, *i.e.*, $h_{GE} = X_{GE}\beta_{GE}$ with $\beta_{GE} \sim N(0,\ \tau_{GE}I)$, which is equivalent to setting $S_{GE} = X_{GE}X_{GE}^T$ (*i.e.*, a linear kernel with $w_m = 1$).

For SimReg, we used the weighted IBS kernel with weight $w_m = (1-q_m)^{24}$. For GESAT, we used R code provided by the authors with the default settings to perform G×E tests (the

Table 1 Type I error rates of SimReg tests with *vs.* without low-rank approximation in rare-variant (RV) simulations

% variance retained in S_G (denoted by p)	Case–control sampling	Random sampling
Joint test $(a_G, a_{GE}) = (0, 0)$		
p = 100%	0.052 (0.0070)[a]	*0.025* (0.0049)
p = 99%	0.052 (0.0070)	0.052 (0.0070)
$G{\times}E$ test $(a_G, a_{GE}) = (0, 0)$		
p = 100%	*0.036* (0.0059)	*0.024* (0.0048)
p = 99%	0.047 (0.0067)	0.043 (0.0064)
$G{\times}E$ test $(a_G, a_{GE}) = (0.02, 0)$		
p = 99% with 20 causal G SNPs	0.064 (0.0077)	0.046 (0.0066)
p = 99% with 40 causal G SNPs	0.049 (0.0068)	0.046 (0.0066)
p = 99% with 60 causal G SNPs	0.045 (0.0066)	0.043 (0.0064)
p = 99% with 80 causal G SNPs	0.050 (0.0069)	0.042 (0.0063)
p = 99% with 100 causal G SNPs	0.041 (0.0063)	0.045 (0.0066)

The corresponding standard errors (SEs) are shown in parentheses. The values in italics are those whose 95% confidence intervals (*i.e.*, rate ± 1.96 × SE) fall below the nominal level. The results are based on 1000 replications. a_G and a_{GE} are the group PARs of the genetic main effect and the $G{\times}E$ effect, respectively.

[a] Standard errors of the type I error rates.

code does not support joint tests). For the burden-based $G{\times}E$ test, we first summarize the marker-set information of subject i, using the number of rare variants in the set, referred to as mutation burden. Then, we fit a logistic model, logit $P(Y_i = 1|X_i, G) = \beta_0 + X_{Ei}\beta_E + \widetilde{G}_i\beta_G + \widetilde{G}_iX_{Ei}\beta_{GE}$, where $\widetilde{G}_i$ is the mutation burden for subject i. Under this model, the $G{\times}E$ effect can be detected by testing $H_0: \beta_{GE} = 0$, and the joint effect can be detected by testing $H_0: \beta_G = \beta_{GE} = 0$.

CV simulations: We obtained 234 phased haplotypes of gene *TCF7L2* from chromosome 10 of the Utah residents with ancestry from northern and western Europe (CEU) samples in HapMap 3. We focused our analysis on the 29 typed SNPs genotyped in the Wellcome Trust Case Control Consortium (WTCCC) analysis (Wellcome Trust Case Control Consortium 2007). The MAFs of these 29 SNPs ranged from 0.0085 to 0.48. We randomly drew 2 haplotypes with replacement from the 234 phased haplotypes to form an individual genotype. We assumed that 2 of the 29 SNPs were causal and simulated the binary phenotype of individual i from a Bernoulli (π_i) distribution, where $\pi_i = e^{\eta_i}/(1 + e^{\eta_i})$, $\eta_i = \gamma_0 + X_{Ei}\gamma_E + G_i^1\gamma_G^1 + G_i^2\gamma_G^2 + G_{1i}X_{Ei}\gamma_{GE}^1 + G_{2i}X_{Ei}\gamma_{GE}^2$, and G_i^r is the number of minor alleles at the causal locus $r = 1, 2$. We generated the ith individual's environmental covariate, X_{Ei}, from a $N(0, 6)$ distribution and set $\gamma_0 = -2.5$, $\gamma_E = \log(1.5) = 0.4055$. As in the RV simulations, we considered case–control sampling (with 750 cases and 750 controls) and random sampling (with sample size 1500 and prevalence rate 0.3).

In the type I error analysis, we set $\gamma_{GE} = \gamma_G^1 = \gamma_G^2 = 0$ for the joint test. For the $G{\times}E$ test, we set $\gamma_{GE} = 0$ and considered $\gamma_G^1 = \gamma_G^2 = 0$ and $\gamma_G^1 = \gamma_G^2 = 1/2 \times \log(1.2) = 0.0912$. We considered five pairs of causal SNPs (*i.e.*, $\gamma_G^r > 0$) with different MAFs as shown in Table 3. We performed 1000 replicates per scenario. In the power analysis, we set $\gamma_G^1 = \gamma_G^2 = 1/2 \times \log(1.2) = 0.0912$ and $\gamma_{GE}^1 = \gamma_{GE}^2 = 1/2 \times \log(1.055) = 0.0268$ for both the $G{\times}E$ test and the joint test. We considered all possible pairs of causal SNPs for a total of $\binom{29}{2} = 406$ scenarios. We performed 100 replicates per scenario. To mimic the typical CV analysis, we excluded the 2 causal SNPs and analyzed the other 27 SNPs only in the association tests. For SimReg, we set the locus-specific weight $w_m = 1$. We compared the proposed SimReg method to GESAT and the single-SNP minimum P-value method (referred to as min-P). For the min-P method, we fitted the model logit $P(Y_i = 1|X_i, G) = \delta_0 + X_{Ei}\delta_E + G_i^m\delta_G + G_i^mX_{Ei}\delta_{GE}$ for each SNP m to obtain the P-values of the $G{\times}E$ test (*i.e.*, testing $H_0: \delta_{GE} = 0$) and the joint test (*i.e.*, testing $H_0: \delta_{GE} = \delta_G = 0$). For a given test (*e.g.*, the $G{\times}E$ test), we took the minimum of the 27 $G{\times}E$ P-values and calculated the adjusted P-value as $1 - \{1 - \min\ P\text{-value}\}^{k_{\text{eff}}}$, where k_{eff} is the effective number of independent tests obtained using the method of Moskvina and Schmidt (2008).

Results and Discussion

Simulation studies

Results of type I error analyses (Table 1, Table 2, and Table 3): The type I errors for the $G{\times}E$ test and the joint test are shown in Table 1 and Table 2 for RV simulations and Table 3 for CV simulations. From Table 1, we see that SimReg can have conservative type I errors when using $P = 100\%$, which can be alleviated by using $P = 99\%$. Table 2 shows that SimReg, burden-based, and GESAT methods all have type I error rates around the nominal level in RV analyses. Table 3 shows that SimReg, min-P, and GESAT all have type I error rates around the nominal level in the CV analyses.

Results of RV power analyses (Figure 1): The power results for a main-effect group PAR (a_G) of 0.02 and a $G{\times}E$ group PAR (a_{GE}) of 0.1 are shown in Figure 1. For the $G{\times}E$ tests and the joint tests, SimReg has higher power than the burden-based test and GESAT ($G{\times}E$ test only) across different numbers of causal SNPs and different study designs. GESAT has the lowest power for the $G{\times}E$ test. Because we assumed a linear $G{\times}E$ effect in the simulation, the power loss may be attributable to the unweighted similarity (*i.e.*, $w_m = 1$), which resulted in an overall similarity score dominated by less-frequent over rare variants and led to little variations among individual pairs.

We note that for both the SimReg and burden-based tests, the power of the joint test is slightly less than the power of the $G{\times}E$ test. It is likely that this is caused by the weak main-effect signal in the simulation: the majority of the simulated data sets had significant $G{\times}E$ effects but negligible genetic main effects. Consequently, compared to the $G{\times}E$ test statistic, the joint test statistic may have incorporated additional noise from the G test statistic, which can result in power loss. We also observe that the power loss in the joint test appears to be larger for SimReg than for the burden-based tests because the degrees of freedom (d.f.) of a SimReg test spent on the G effect tend to be higher than those of a burden-based test. However, the power of SimReg is still higher than that of the burden-based test, and the additional d.f. consumed by SimReg (compared to the burden-based test) ensure robustness against between-locus etiological heterogeneities (Pongpanich *et al.* 2012) as well as against model misspecifications.

Table 2 Type I error rates of the $G{\times}E$ test and the joint test for rare-variant (RV) simulations

Nominal level	SimReg[a]	Burden-based	GESAT
Joint test $(a_G, a_{GE}) = (0, 0)$			
0.05	0.0504 (0.0022)[b]	0.0511 (0.0022)	NA
0.01	0.0093 (0.0010)	0.0110 (0.0010)	NA
0.005	0.0047 (0.0007)	0.0056 (0.0007)	NA
0.001	0.0010 (0.0003)	0.0011 (0.0003)	NA
$G{\times}E$ test $(a_G, a_{GE}) = (0, 0)$			
0.05	0.0496 (0.0022)	0.0523 (0.0022)	**0.05090** (0.0024)
0.01	0.0085 (0.0009)	0.0104 (0.0010)	0.0119 (0.0011)
0.005	*0.0038* (0.0006)	0.0044 (0.0007)	0.0050 (0.0007)
0.001	0.0007 (0.0026)	0.0008 (0.0003)	0.0007 (0.0003)
$G{\times}E$ test $(a_G, a_{GE}) = (0.02, 0)$[c]			
0.05	0.0473 (0.0021)	0.0482 (0.0021)	**0.0602** (0.0024)
0.01	0.0099 (0.0010)	0.0112 (0.0011)	0.0119 (0.0011)
0.005	0.0052 (0.0007)	0.0055 (0.0007)	0.0062 (0.0008)
0.001	0.0014 (0.0004)	0.0010 (0.0003)	0.0009 (0.0003)

Data were generated using a case–control design. The corresponding standard errors (SEs) are shown in parentheses. The values in italics/boldface type are those whose 95% confidence intervals (*i.e.*, rate ± 1.96 × SE) fall below/above the nominal level. a_G and a_{GE} are the group PARs of the genetic main effect and the $G{\times}E$ effect, respectively. The results were obtained based on 10,000 replications.

[a] Using p (the proportion of variation explained by the leading eigenvalues in S_G) = 0.99.

[b] Standard errors of the type I error rates.

[c] Assuming 40 SNPs with causal main (G) effect.

Results of CV power analyses (Figure 2): To present the power results of the $\binom{29}{2} = 406$ scenarios, we grouped the scenarios into three categories based on the LD structure between the causal SNPs and the analyzed SNPs. The three LD groups, *i.e.*, the lower one-third (low LD), the middle one-third (medium LD), and the top one-third (high LD), are defined based on the average of 54 R^2 values, where each value is the R^2 between a causal SNP (2 in total) and an analyzed SNP (27 in total). We present side-by-side boxplots of the power of SimReg, min-P, and GESAT (for $G{\times}E$ tests) as well as the mean power value in Figure 2. We observe that when the LD is lower, the power of all methods is lower. This is expected because under low-LD scenarios the markers contain less information about the 2 causal loci. For the $G{\times}E$ test (Figure 2, top), SimReg and GESAT have very similar power, as expected because both methods set $w_m = 1$. The powers of SimReg and min-P are similar when LD is low. As the LD increases, SimReg starts to have power improvement over min-P. The difference becomes more obvious when LD is high. For the joint test (Figure 2, bottom), the relative power of SimReg *vs.* min-P is similar to what was observed for the $G{\times}E$ tests. Furthermore, the relative performance between SimReg and min-P for binary traits is similar to what was observed for quantitative traits (Tzeng *et al.* 2011).

Data Applications

Analysis of gene-by-physical activity effect on obesity, using CoLaus samples: We used Sanger sequence data of the *PLA2G7* gene for 1961 subjects from the CoLaus (Song *et al.* 2012) and studied PLA2G7's association with the levels of lipoprotein-associated phospholipase A2 (Lp-PLA2). The CoLaus study of Firmann *et al.* (2008) is a population-based study to assess the risk factors of cardiovascular disease (CVD) in Caucasian residents of Lausanne, Switzerland aged 35–75 years. *PLA2G7* encodes Lp-PLA2, and the elevated plasma levels of Lp-PLA2 activity have been shown to be associated with increased risk of coronary heart disease (Thompson *et al.* 2010). We imputed sporadic missing genotypes, using the MaCH software package (Li *et al.* 2010), and obtained a total of 100 SNPs with MAF < 0.05 (range from 0.000255 to 0.029).

The genetic influence of *PLA2G7* on the body mass affected by exercise has been reported in the literature (Wootton *et al.* 2007; Detopoulou *et al.* 2009). The potential modulating effect of *PLA2G7* on arachidonic acid was hypothesized to be related to the association between the *PLA2G7* variants and a reduced risk of coronary artery disease (Ninio *et al.* 2004; Wootton *et al.* 2007). Using *PLA2G7* as a positive control, we investigated the potential interaction between physical activity and genetic variants on BMI. We defined obesity as BMI > 30 and evaluated the effects of *PLA2G7* (G), physical activity (E), and $G{\times}E$ interactions on obesity. We considered three methods: SimReg, GESAT, and the burden-based test. In all analyses, we adjusted for age, sex, ethnic background (five PCs), smoking status, and alcohol consumption. For SimReg, we used weight $w_m = (1 - q_m)^{24}$ and a low-rank approximation with $\mathrm{p} = 0.99$; the resulting P-values of the joint test and the $G{\times}E$ test were 1.46×10^{-3} and 1.05×10^{-3}, respectively, which suggested that *PLA2G7* may affect the influence of physical activity on obesity. GESAT, which set

Table 3 Type I error rates of the *G*×*E* test and the joint test for common-variant (CV) simulations

Effect size considered	MAFs of the causal SNPs	SimReg	min-*P*	GESAT
Joint test ($\gamma_G^r = \gamma_{GE}^r = 0$)	NA	0.044 (0.0065)[a]	0.060 (0.0075)	NA
G×*E* test ($\gamma_{GE}^r = 0$)				
$\gamma_G^r = 0$	NA	*0.037* (0.0060)	0.054 (0.0072)	*0.036* (0.0059)
$\gamma_G^r = 0.0912$	0.009, 0.094	0.042 (0.0068)	0.040 (0.0062)	0.053 (0.0071)
$\gamma_G^r = 0.0912$	0.009, 0.1966	0.040 (0.0062)	0.043 (0.0064)	0.055 (0.0072)
$\gamma_G^r = 0.0912$	0.094, 0.1966	0.040 (0.0062)	0.045 (0.0066)	**0.070** (0.0081)
$\gamma_G^r = 0.0912$	0.1966, 0.2222	0.047 (0.0067)	0.044 (0.0065)	0.051 (0.0070)
$\gamma_G^r = 0.0912$	0.2991, 0.4188	0.049 (0.0068)	0.050 (0.0069)	0.054 (0.0072)

The corresponding standard errors (SEs) are shown in parentheses. The values in italics/boldface type are those whose 95% confidence intervals (*i.e.*, rate ± 1.96 × SE) fall below/above the nominal level. The results were obtained based on 1000 replications, and γ_G^r and γ_{GE}^r are the effect sizes of the causal SNPs for the main effect and the *G*×*E* effect, respectively.

[a] Standard errors of the type I error rates.

$w_m = 1$, yielded a *G*×*E* *P*-value of 0.637. These results are not unexpected given the simulation results; *i.e.*, the unweighted similarity scores did not have power to detect rare variants because the contribution from rarer variants may be overwhelmed by the less rare variants during collapsing. The *P*-values of the burden-based tests were 0.013 for the joint test and 3.84×10^{-3} for the *G*×*E* test, which are larger than SimReg *P*-values but give the same significant conclusions as SimReg. The results agree with the observation from the RV simulations that the proposed method is more powerful in detecting *G*×*E* effects.

Analysis of TCF7L2-by-BMI effect on type 2 diabetes, using WTCCC samples: The data were obtained from the type 2 diabetes (T2D) case–control study conducted by the WTCCC (Wellcome Trust Case Control Consortium 2007). The controls were samples from the 1958 British Birth Cohort. The case samples were collected from various sites across the United Kingdom to be comparable to the controls. The genotyping was conducted on an Affymetrix 500K chip. Previous genome-wide association studies (Timpson *et al.* 2009) have indicated an interaction between TCF7L2 and BMI on T2D. Treating this TCF7L2 × BMI effect on T2D as a true positive, we evaluated the performance of the proposed SimReg test (with weight $w_m = 1$) and compared to GESAT and the min-*P* test.

We fitted a model where the response variable is the T2D status and the explanatory variables include the 29 SNPs in TCF7L2, BMI, TCF7L2×BMI, and sex. After applying sample and SNP quality control filters to remove substantial missing data, the data set contained 1913 cases and 1455 controls. We first performed the joint test and obtained a *P*-value of 1.81×10^{-10} for SimReg and 1.39×10^{-9} for min-*P*. The gene-level *P*-value of min-*P* is obtained as $1 - (1 - \min_{1 \le \ell \le 29} P\text{-val}_\ell)^{K_{\text{eff}}}$, where $K_{\text{eff}} = 19.8$ is the effective number of independent tests for TCF7L2 estimated by Moskvina and Schmidt (2008). The *P*-values of the *G*×*E* tests are 4.05×10^{-5} for SimReg, 6.74×10^{-6} for GESAT, and 2.72×10^{-3} for min-*P* (adjusted *P*-value). The difference between SimReg and GESAT *P*-values can be attributed to the different choices of kernels (*e.g.*, IBS kernel for SimReg *vs.* linear kernel for GESAT) and the different algorithm to estimate the nuisance main effects (*e.g.*, EM algorithm *vs.* ridge penalization). The relatively large *P*-values of min-*P* suggest that there may be multiple moderate-effect loci in TCF7L2 contributing to the T2D risk, as opposed to a few strong-effect loci. The magnitude of the *P*-value difference in the joint tests was relatively small compared to the *P*-value difference in *G*×*E* tests, suggesting a strong main effect of TCF7L2 on T2D as shown in the literature (Helgason *et al.* 2007; Scott *et al.* 2007).

Conclusion

In this article we proposed a marker-set method based on similarity regression to examine *G*×*E* effects for binary traits and showed it is computationally feasible, powerful, and applicable to both common and rare variants. By demonstrating the equivalence of our gene-similarity regression model to a GLMM framework, we showed that SimReg is robust against model misspecification, like other random-effects-based approaches (*e.g.*, Lin *et al.* 2013). However, because the structure of S_{GE} is atypical, one cannot apply the general score test of GLMM as implemented in existing statistical software because it often yields invalid estimates of τ_G (*e.g.*, negative values). We developed an EM algorithm to address the challenges associated with estimation and computation encountered in GLMM model fitting. The C code that implements the proposed joint and *G*×*E* tests is available at http://www4.stat.ncsu.edu/~jytzeng/software_simreg.php. We demonstrated the utility of SimReg in rare variant *G*×*E* analysis. We also found that for RVs, the low-rank approximation to the main-effect similarity matrix (S_G) is necessary to avoid an overconservative type I error rate.

One possible strategy to apply the proposed SimReg tests is to start with a joint test to detect the overall association induced by the *G* main effect or the *G*×*E* effects. A screening by joint tests may lead to increased flexibility and power to detect a signal because some genes can exhibit negligible marginal effects but strong effects among particular exposure groups (Kraft *et al.* 2007; Thomas 2010). If the joint test is rejected, a *G*×*E* test can then be used to identify whether the effects of the genetic variables are modified by the environmental variables.

One can view the SimReg framework as an implementation of a class of models for modeling h_{GE}, which includes GESAT as a special case. In SimReg, one can determine how

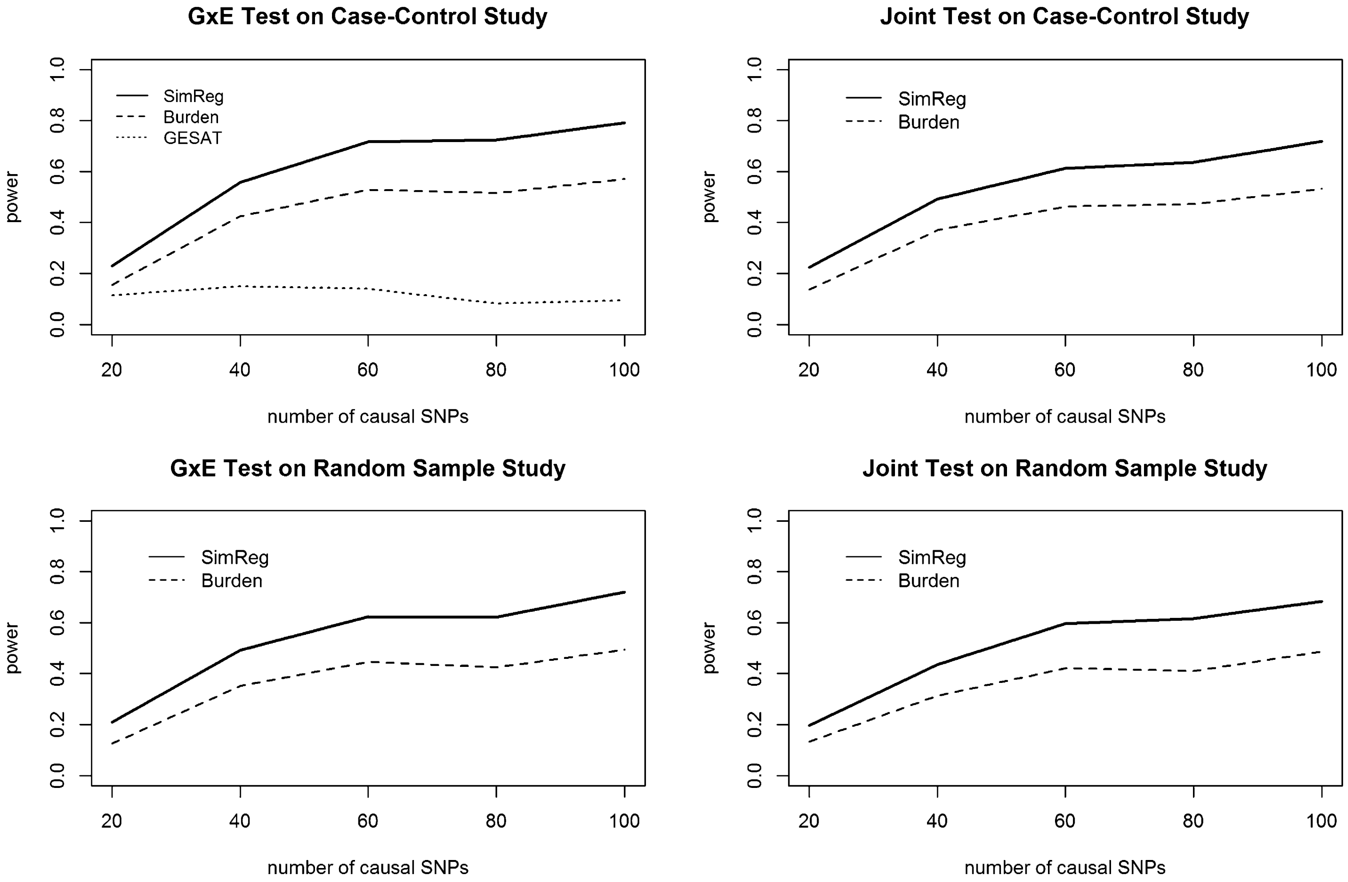

Figure 1 Power of $G{\times}E$ and joint tests for rare-variant simulations. The powers of SimReg, burden-based, and GESAT tests are represented by the solid (–), dashed (- - -), and dotted (⋯) lines, respectively. GESAT is performed only under case–control studies. The results were obtained based on 500 replicates.

the $G{\times}E$ effect is modeled by specifying a certain similarity metric, *e.g.*, linear kernel, IBS kernel, or quadratic kernel, as well as by imposing variant-specific weights when collapsing the information across markers. If a linear kernel is used with $w_m = 1$, the SimReg $G{\times}E$ test is equivalent to GESAT. However, one subtle difference is that SimReg uses an EM algorithm to estimate the nuisance main effects, whereas GESAT uses a penalized method. Another remark concerns the role of the variant-specific weight based on MAFs. As we observed in the numerical studies, although the unweighted similarity performed satisfactorily in CV analyses, it has little power in RV analyses. This is because the sum of unweighted similarity scores would be dominated by information from nonrare events. Consequently, when rare variants are studied, the multimarker similarity scores would exhibit little variation. The MAF-based weights in essence perform a soft thresholding to downweight or diminish the contribution of less-frequent or common variants in the multimarker similarity score.

The rationale of a collapsing analysis is to detect the amplified effects of rare variants in aggregate. Experience from main-effect testing suggests that variance component-based tests such as SimReg would have better power than burden-based tests if genetic effects vary radically across variants or if many null variants exist in the set (Pongpanich *et al.* 2012; Lee *et al.* 2014). However, the presence of many null variants can still unfavorably affect the test performance. For main-effect collapsing tests, efforts have been made to boost power when the signal sparsity is low by adaptively focusing on the subsets enriched with causal variants (*e.g.*, Barnett 2014; Pan *et al.* 2014). Their extensions to $G{\times}E$ tests will be helpful to further optimize the power to detect $G{\times}E$ effects.

In this work we focused on examining the $G{\times}E$ interaction effect for a single environmental factor. However, a similar model involving multiple $G{\times}E$ interaction effects could be fitted. This method could be easily extended to test for gene–gene interaction in cases where one gene is suspected to interplay with other genes.

Acknowledgments

The authors are grateful to Mark McCarthy, Timothy Frayling, William Rayner, and members of the Warren 2 Consortium for providing the BMI data. This work makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. The authors are also grateful to Peter Vollenweider

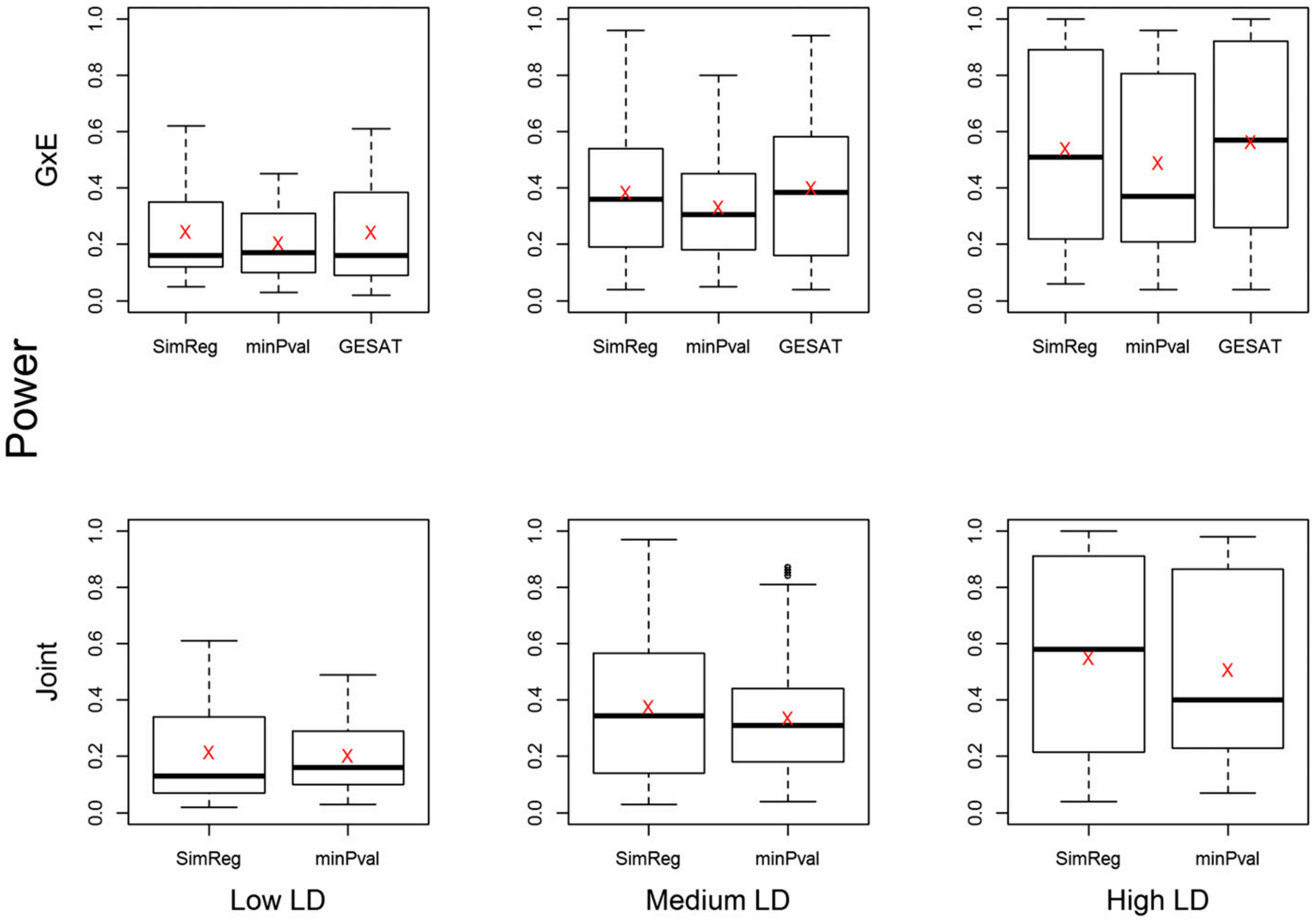

Figure 2 Power of $G \times E$ and joint tests for common-variant simulations. The side-by-side boxplots show the powers of the proposed SimReg method, the minimum *P*-value method, and GESAT. A total of $\binom{29}{2} = 406$ scenarios were considered (*i.e.*, letting each SNP pair of the 29 SNPs be causal), with 100 replicates per scenario. The 406 scenarios were classified into three groups based on the LD pattern between the 2 causal SNPs and the remaining 27 SNPs. The red "X" in each boxplot represents the average power for each LD group.

and Gerard Waeber, Principal Investigators of the CoLaus study, and Meg Ehm and Matthew Nelson, collaborators at GlaxoSmithKline, for providing the CoLaus phenotype and sequence data. The authors thank Michael Wu for helping with the COSI simulation design and thank Shannon Holloway for her constructive input to improve the manuscript. The CoLaus study is supported by research grants from GlaxoSmithKline, the Faculty of Biology and Medicine of Lausanne, Switzerland, and by the Swiss National Science Foundation grants 33CSCO-122661 and 33CS30-139468. This work was partially supported by National Institutes of Health grants R01 MH084022 (to G.Z., D.Z., and J.-Y.T.), T32GM081057 (to R.M.), R01 CA85848-12 (to D.Z.), and P01 CA142538 (to J.-Y.T.).

Literature Cited

Barnett, I. J., 2014 SNP-set tests for sequencing and genome-wide association studies. Ph.D. Dissertation, Harvard University, Cambridge, MA. Available at: http://nrs.harvard.edu/urn-3:HUL.InstRepos:12274530

Beckmann, L., C. Fischer, M. Obreiter, M. Rabes, and J. Chang-Claude, 2005 Haplotype-sharing analysis using Mantel statistics for combined genetic effects. BMC Genet. 6(Suppl. 1): S70.

Cai, T., G. Tonini, and X. Lin, 2011 Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67: 975–986.

Chatterjee, N., Z. Kalaylioglu, R. Moslehi, U. Peters, and S. Wacholder, 2006 Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am. J. Hum. Genet. 79(6): 1002–1016.

Dai, J. Y., B. A. Logsdon, Y. Huang, L. Hsu, A. P. Reiner *et al.*, 2012 Simultaneously testing for marginal genetic association and gene-environment interaction. Am. J. Epidemiol. 176(2): 164–173.

Detopoulou, P., T. Nomikos, E. Fragopoulou, D. B. Panagiotakos, C. Pitsavos *et al.*, 2009 Lipoprotein-associated phospholipase A2 (Lp-PLA2) activity, platelet-activating factor acetylhydrolase (PAF-AH) in leukocytes and body composition in healthy adults. Lipids Health Dis. 8: 19.

Duchesne, P., and P. Lafaye De Micheaux, 2010 Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. Comput. Stat. Data Anal. 54(4): 858–862.

Elston, R., C. S. Buxbaum, K. B. Jacobs, and J. M. Olson, 2000 Haseman and Elston revisited. Genet. Epidemiol. 19: 1–17.

Fan, R., and S. H. Lo, 2013 A robust model-free approach for rare variants association studies incorporating gene-gene and gene-environmental interactions. PLoS ONE 8(12): e83057.

Firmann, M., V. Mayor, P. Vidal, M. Bochud, A. Pecoud *et al.*, 2008 The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc. Disord. 8(1): 6.

Haseman, J. K., and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. 2(1): 3–19.

Helgason, A., S. Pálsson, G. Thorleifsson, S. F. A. Grant, V. Emilsson *et al.*, 2007 Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. Nat. Genet. 39: 218–225.

Jiao, S., L. Hsu, S. Bézieau, H. Brenner, A. T. Chan *et al.*, 2013 SBERIA: set based gene environment interaction test for rare and common variants in complex diseases. Genet. Epidemiol. 37: 452–464.

Kraft, P., Y. C. Yen, D. O. Stram, J. Morrison, and W. J. Gauderman, 2007 Exploiting gene-environment interaction to detect genetic associations. Hum. Hered. 63(2): 111–119.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, 2010 MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34(8): 816–834.

Lin, X., S. Lee, D. C. Christiani, and X. Lin, 2013 Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics 14: 667–681.

Lee, S., G. R. Abecasis, M. Boehnke, and X. Lin, 2014 Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. 95: 5–23.

Madsen, B. E., and S. R. Browning, 2009 A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5: e1000384.

Miao, H., 2009 Model selection and estimation in additive regression models. Ph.D. Dissertation, North Carolina State University, Raleigh, NC.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753.

Mechanic, L. E., H.-S. Chen, C. I. Amos, N. Chatterjee, N. J. Cox *et al.*, 2012 Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. Genet. Epidemiol. 36: 22–35.

Moskvina, V., and K. M. Schmidt, 2008 On multiple-testing correction in genome-wide association studies. Genet. Epidemiol. 32: 567–573.

Mukherjee, B., and N. Chatterjee, 2008 Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics. 64:685–94.

Murcray, C. E., J. P. Lewinger, and W. J. Gauderman, 2009 Gene-environment interaction in genome-wide association studies. Am. J. Epidemiol. 169(2): 219–226.

Ninio, E., D. Tregouet, J. L. Carrier, D. Stengel, C. Bickel *et al.*, 2004 Platelet-activating factor-acetylhydrolase (PAF-AH) and PAF-receptor gene haplotypes in relation to future cardiovascular events in patients with coronary artery disease. Hum. Mol. Genet. 13(13): 1341–1351.

Pan, W., J. Kim, Y. Zhang, X. Shen, and P. Wei, 2014 A powerful and adaptive association test for rare variants. Genetics 197: 1081–1095.

Price, A. L., G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples *et al.*, 2010 Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. 86(6): 832–838.

Pongpanich, M., M. L. Neely, and J.-Y. Tzeng, 2012 On the aggregation of multimarker information for marker-set and sequencing data analysis: genotype collapsing *vs.* similarity collapsing. Front. Genet. 2: 1–14.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15(11): 1576–1583.

Schaid, D. J., 2010a Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum. Hered. 70(2): 109–131.

Schaid, D. J., 2010b Genomic similarity and kernel methods II: genomic information. Hum. Hered. 70(2): 132–140.

Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li *et al.*, 2007 A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341–1345.

Sham, P., and S. Cherny, 2011 *Analysis of Complex Disease Association Studies [Electronic Re-Source]: A Practical Guide*. Academic Press/Elsevier, London/Burlington, MA.

Sohns, M., E. Viktorova, C. I. Amos, and P. Brennan, G. Fehringer *et al.*, 2013 Empirical hierarchical Bayes approach to gene–environment interactions: development and application to genome–wide association studies of lung cancer in TRICL. Genet. Epidemiol. 37: 551–559.

Song, K., M. R. Nelson, J. Aponte, E. S. Manas, S. A. Bacanu *et al.*, 2012 Sequencing of Lp-PLA2-encoding *PLA2G7* gene in 2000 Europeans reveals several rare loss-of-function mutations. Pharmacogenomics J. 12(5): 425–431.

Thomas, D., 2010 Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu. Rev. Public Health 31: 21–36.

Thomas, D., 2011 Response to 'Gene-by-environment experiments: a new approach to finding the missing heritability' by Van Ijzendoorn *et al.* Nat. Rev. Genet. 12(12): 881.

Thompson, A., P. Gao, L. Orfei, S. Watson, A. E. Di *et al.*, 2010 Lipoprotein-associated phospholipase A2 and risk of coronary disease, stroke, and mortality: collaborative analysis of 32 prospective studies. Lancet 375: 1536–1544.

Timpson, N. J., C. M. Lindgren, M. N. Weedon, J. Randall, W. H. Ouwehand *et al.*, 2009 Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. Diabetes 58: 505–510.

Tzeng, J.-Y., B. Devlin, L. Wasserman, and K. Roeder, 2003 On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am. J. Hum. Genet. 72(4): 891–902.

Tzeng, J.-Y., and D. Zhang, 2007 Haplotype-based association analysis via variance component score test. Am. J. Hum. Genet. 81: 927–938.

Tzeng, J.-Y., D. Zhang, S.-M. Chang, D. C. Thomas, and M. Davidian, 2009 Gene-trait similarity regression for multimarker-based association analysis. Biometrics 65: 822–832.

Tzeng, J.-Y., D. Zhang, M. Pongpanich, C. Smith, M. I. McCarthy *et al.*, 2011 Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am. J. Hum. Genet. 89(2): 277–288.

van Os, J., and B. Rutten, 2009 Gene-environment-wide interaction studies in psychiatry. Am. J. Psychiatry 166(9): 964–966.

Voorman, A., T. Lumley, B. McKnight, and K. Rice, 2011 Behavior of qq-plots and genomic control in studies of gene-environment interaction. PLoS ONE 6: e19416.

Wang, T., and R. C. Elston, 2005 Two-level Haseman-Elston regression for general pedigree data analysis. Genet. Epidemiol. 29: 12–22.

Wang, X., N. J. Morris, X. Zhu, and R. C. Elston, 2013 A variance component based multi-marker association test using family and unrelated data. BMC Genet. 14(1): 1–8.

Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.

Wessel, J., and N. J. Schork, 2006 Generalized genomic distance-based regression methodology for multilocus association analysis. Am. J. Hum. Genet. 79: 792–806.

Winham, S. J., and J. M. Biernacka, 2013 Gene–environment interactions in genome–wide association studies: current approaches and new directions. J. Child Psychol. Psychiatry 54(10): 1120–1134.

Wootton, P. T., D. M. Flavell, H. E. Montgomery, M. World, S. E. Humphries *et al.*, 2007 Lipoprotein-associated phospholipase A2 A379V variant is associated with body composition changes in response to exercise training. Nutr. Metab. Cardiovasc. Dis. 17(1): 24–31.

Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock *et al.*, 2010 Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. 86: 929–942.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare variant association testing for sequencing data using the sequence kernel association test (skat). Am. J. Hum. Genet. 89: 82–93.

Zhang, D., and X. Lin, 2003 Hypothesis testing in semiparametric addictive mixed models. Biostatistics 4: 57–74.

Communicating editor: I. Hoeschele

Appendix A: Marginal Trait Covariance cov(Y_i, Y_j)

Define $h_i = h_{Gi} + h_{GEi}$. Under GLMM (2),

$$
\begin{aligned}
&\operatorname{cov}\left(Y_i, Y_j\right)\\
&= \operatorname{cov}_h\left\{E(Y_i \middle| X, h), E(Y_j \middle| X, h)\right\} + E_h\left\{\operatorname{cov}(Y_i, Y_j \middle| X, h)\right\}\\
&= \operatorname{cov}_h\left\{E(Y_i \middle| X, h), E(Y_j \middle| X, h)\right\}\\
&\quad (\because \text{ conditional independence of } Y_i \text{ and } Y_j)\\
&= \operatorname{cov}_h\left\{g^{-1}(X_i\gamma + h_i), g^{-1}\left(X_j\gamma + h_j\right)\right\}\\
&\approx \operatorname{cov}_h\left\{\begin{array}{l}\left[g^{-1}(X_i\gamma + Eh_i) + \left[\left.\dfrac{\partial g^{-1}(X_i\gamma + h_i)}{\partial h_i}\right|_{h_i = Eh_i}\right](h_i - Eh_i)\right],\\ \left[g^{-1}\left(X_j\gamma + Eh_j\right) + \left[\left.\dfrac{\partial g^{-1}\left(X_j\gamma + h_j\right)}{\partial h_j}\right|_{h_j = Eh_j}\right]\left(h_j - Eh_j\right)\right]\end{array}\right\}
\end{aligned}
$$

[by taking the first-order Taylor expansion of $g^{-1}(X_i\gamma + h_i)$ with respect to h_i around $Eh_i = 0$]

$$
\begin{aligned}
&= \operatorname{cov}_h\left\{\begin{array}{l}\left[g^{-1}(X_i\gamma) + \left[\left.\dfrac{\partial g^{-1}(X_i\gamma + h_i)}{\partial h_i}\right|_{h_i = 0}\right] \times h_i\right],\\ \left[g^{-1}\left(X_j\gamma\right) + \left[\left.\dfrac{\partial g^{-1}\left(X_j\gamma + h_j\right)}{\partial h_j}\right|_{h_j = 0}\right] \times h_j\right]\end{array}\right\}\\
&= \left[\left.\frac{\partial g^{-1}(X_i\gamma + h_i)}{\partial h_i}\right|_{h_i = 0}\right] \times \left[\left.\frac{\partial g^{-1}\left(X_j\gamma + h_j\right)}{\partial h_j}\right|_{h_j = 0}\right] \times \operatorname{cov}\left(h_i, h_j\right)\\
&= \left[\left.\frac{\partial g^{-1}(X_i\gamma + h_i)}{\partial h_i}\right|_{h_i = 0}\right] \times \left[\left.\frac{\partial g^{-1}\left(X_j\gamma + h_j\right)}{\partial h_j}\right|_{h_j = 0}\right] \times \operatorname{cov}(h_{Gi} + h_{GEi}, h_{Gj} + h_{GEj})\\
&= \left\{\frac{\partial g(\mu_i^0)}{\mu_i^0}\right\}^{-1} \times \left\{\frac{\partial g(\mu_j^0)}{\mu_j^0}\right\}^{-1} \times \left\{cov(h_{Gi}, h_{Gj}) + cov(h_{GEi}, h_{GEj})\right\}\\
&= \left\{g'(\mu_i^0) g'\left(\mu_j^0\right)\right\}^{-1} \times \left\{\tau_G S_{ij} + \tau_{GE} X_{Ei} X_{Ej} S_{ij}\right\},
\end{aligned}
$$

Appendix B: EM Algorithm to Estimate τ_G and σ in the SimReg $G \times E$ Test

Under the null hypothesis $H_0^{GE}: \tau_{GE} = 0$, model (3) becomes $g(\mu) = X\gamma + Z_G b$ with $b \sim N(0, \tau_G I_{L \times L}\ )$. Let $Y = (Y_1, \ldots, Y_n)$ be the vector of binary traits, and let $\theta = (\gamma, \tau_G)$ be the parameter vector. We consider an expectation-maximization algorithm based on observed data Y and missing data b. Let $\log f(Y, b; \theta)$ be the complete data log-likelihood. In the expectation step (E-step), we compute $Q(\theta \middle| \theta^{(t)})$ as

$$
\begin{aligned}
Q(\theta \middle| \theta^{(t)}) &= E\left\{\log f(Y, b; \theta) \middle| Y, \theta^{(t)}\right\}\\
&= E\left\{\log f(Y \middle| b; \theta) \middle| Y, \theta^{(t)}\right\} + E\left\{\log f(b; \theta) \middle| Y, \theta^{(t)}\right\},
\end{aligned}
$$

because $f(Y, b; \theta) = f(Y|b; \theta) f(b; \theta)$. For the first term, we have

$$E\Big\{\log f(Y|b;\theta)|Y;\theta^{(t)}\Big\} = \sum_{i=1}^{n} E\Big\{Y_i \log \mu_i + (1-Y_i)\log(1-\mu_i)|Y,\theta^{(t)}\Big\}. \tag{B1}$$

For the second term, note that

$$\begin{aligned}
\log f(b;\theta) &= \log f(b;\tau_G)\\
&= \log\Big\{(2\pi)^{-(L/2)}\big|\tau_G I_L\big|^{-(1/2)}\exp\Big\{-\frac{1}{2}b^T(\tau_G I_L)^{-1}b\Big\}\Big\}\\
&= -\frac{L}{2}\log 2\pi - \frac{L}{2}\log \tau_G - \frac{b^Tb}{2\tau_G},
\end{aligned}$$

where $\big|\tau_G I_L\big| = \tau_G^L$. Therefore,

$$\begin{aligned}
E\Big\{\log f(b;\theta)\big|Y,\theta^{(t)}\Big\} &= E\Big\{\Big(-\frac{L}{2}\log 2\pi - \frac{L}{2}\log\tau_G - \frac{b^Tb}{2\tau_G}\Big)\Big|Y,\theta^{(t)}\Big\}\\
&= -\frac{L}{2}\log 2\pi - \frac{L}{2}\log\tau_G - \frac{E\Big(b^Tb|Y,\theta^{(t)}\Big)}{2\tau_G}.
\end{aligned} \tag{B2}$$

By expressing the complete-data log-likelihood in two parts, the fixed effect γ occurs only in the first term $E\{\log f(Y\big|b,\theta^{(t)})\}$ and variance component τ_G occurs only in the second term, $E\{\log f(b;\theta)\big|Y,\theta^{(t)}\}$. Thus, the maximization steps for obtaining $\widehat{\tau_G}^{(t+1)}$ and $\hat{\gamma}^{(t+1)}$ can be discussed separately.

Maximization step for obtaining $\widehat{\tau_G}^{(t+1)}$

To obtain $\widehat{\tau_G}^{(t+1)}$, we can focus on $E\{\log f(b;\theta)\big|Y,\theta^{(t)}\}$ We take the derivative of (B2) with respect to τ_G and get

$$\frac{\partial E\Big\{\log f(b;\theta)\big|Y,\theta^{(t)}\Big\}}{\partial\tau_G} = -\frac{L}{2\tau_G} + \frac{E(b^Tb\big|Y,\theta^{(t)})}{2\tau_G^2}.$$

Setting this equal to zero, we get

$$\begin{aligned}
\widehat{\tau_G}^{(t+1)} &= \frac{E(b^Tb\big|Y,\theta^{(t)})}{L}\\
&= \frac{1}{L}\Big[b^{(t)T}b^{(t)} + \text{trace}\Big(\Sigma^{(t)}\Big)\Big].
\end{aligned} \tag{B3}$$

Equation B3 follows because $(b\big|Y,\theta^{(t)}) \sim N(b^{(t)},\Sigma^{(t)})$ approximately. To derive this approximation, we first reexpress $f(Y,b)$ as $f(Y|b)f(b)$, *i.e.*, a product of a Gaussian kernel and some function of Y. Finally, because $f(Y,b) = f(b|Y)f(Y)$, we have $f(b|Y)\dot{\sim}N$. We provide the details in the next subsection.

Derivation of $f(b|Y)$ as well as its mean $b^{(t)}$ and variance $\Sigma^{(t)}$

$$\begin{aligned}
f(Y,b;\theta^{(t)}) &= f(Y\big|b;\theta^{(t)})f(b;\theta^{(t)})\\
&= \prod_{i=1}^{n}\Big\{\mu_i^{Y_i}(1-\mu_i)^{1-Y_i}(2\pi)^{-(L/2)} - \tau_G^{-(L/2)}\exp\Big(-\frac{b^Tb}{2\tau_G}\Big)\Big\}\\
&= \exp\Big\{\sum_{i=1}^{n}[Y_i\log\mu_i + (1-Y_i)\log(1-\mu_i)] - \frac{L}{2}\log 2\pi - \frac{L}{2}\log\tau_G - \frac{b^Tb}{2\tau_G}\Big\}\\
&= \exp\{h(b)\},
\end{aligned}$$

where

$$h(b) = \sum_{i=1}^{n} [Y_i \log \mu_i + (1 - Y_i) \log(1 - \mu_i)] - \frac{L}{2} \log 2\pi - \frac{L}{2} \log \tau_G - \frac{b^T b}{2\tau_G}. \tag{B4}$$

Let $b^{(t)}$ be the value that maximizes $h(b)$; *i.e.*, $h'(b^{(t)}) = 0$. By a Taylor expansion of $h(b)$ with respect to b around $b^{(t)}$, we have

$$h(b) \approx h(b^{(t)}) + h'(b^{(t)})(b - b^{(t)}) + \frac{1}{2}(b - b^{(t)})^T h'' \,(b^{(t)})(b - b^{(t)}) = h(b^{(t)}) + \frac{1}{2}(b - b^{(t)})^T h''(b^{(t)})(b - b^{(t)}).$$

Therefore, the complete data log-likelihood can be approximated by

$$f(Y, b; \theta^{(t)}) \approx \exp\left\{ h(b^{(t)}) + \frac{1}{2}(b - b^{(t)})^T h'' \,(b^{(t)})(b - b^{(t)}) \right\} = \exp\left\{ h(b^{(t)}) \right\} \exp\left\{ \frac{1}{2}(b - b^{(t)})^T h''(b^{(t)})(b - b^{(t)}) \right\}. \tag{B5}$$

In Equation B5, $\exp\{-(1/2)(b - b^{(t)})^T[-h''(b^{(t)})](b - b^{(t)})\}$ is a Gaussian kernel with $-h''(b^{(t)}) = [\Sigma^{(t)}]^{-1}$. Thus, the conditional distribution of $(b|Y; \theta^{(t)})$ approximately follows a multivariate normal distribution with mean vector $b^{(t)}$ and variance–covariance matrix $\Sigma^{(t)} = [-h''(b^{(t)})]^{-1}$.

Next we calculate $h'(b)$ and $h''(b)$. In Equation B4, we rewrite μ_i as $\mu_i(b)$ to emphasize that it is a function of b; *i.e.*, $\mu_i(b) = \exp(X_i\gamma + Z_i b)/(1 + \exp(X_i\gamma + Z_i b))$ with $Z_{i(1 \times L)}$, the ith row of matrix Z_G. Note that $\dot{\mu}_i(b)_{L \times 1} \equiv \partial\mu_i(b)/\partial b = Z_i^T(\exp(X_i\gamma + Z_i b))/\{1 + \exp(X_i\gamma + Z_i b)\}^2 = Z_i^T \mu_i(b)\{1 - \mu_i(b)\}$. Then

$$\begin{aligned} h'(b) = \frac{\partial h(b)}{\partial b} &= \sum_{i=1}^{n} \left\{ Y_i \times \frac{\dot{\mu}_i(b)}{\mu_i(b)} + (1 - Y_i) \times \frac{-\dot{\mu}_i(b)}{1 - \mu_i(b)} \right\} - \frac{b}{\tau_G} \\ &= \sum_{i=1}^{n} \left\{ Z_i^T Y_i \{1 - \mu_i(b)\} - Z_i^T (1 - Y_i) \mu_i(b) \right\} - \frac{b}{\tau_G} \\ &= \sum_{i=1}^{n} \left\{ Z_i^T Y_i - Z_i^T \mu_i(b) \right\} - \frac{b}{\tau_G} \\ &= Z^T Y - Z^T \mu(b) - \frac{b}{\tau_G} \\ &= Z^T (Y - \mu(b)) - \frac{b}{\tau_G}, \end{aligned}$$

where $\mu(b) = (\mu_1(b), \mu_2(b), \dots, \mu_n(b))^T$, and

$$\begin{aligned} h''\,(b) = \frac{\partial h'(b)}{\partial b^T} &= \sum_{i=1}^{n} \left\{ Z_i^T Y_i - Z_i^T \dot{\mu}_i(b) \right\} - \frac{1}{\tau_G} I_L \\ &= -\sum_{i=1}^{n} \mu_i(b)\{1 - \mu_i(b)\} Z_i^T Z_i - \frac{1}{\tau_G} I_L \\ &= -Z^T W(b) Z - \frac{1}{\tau} I_L \\ &= -(Z^T W(b) Z + \frac{1}{\tau} I_L), \end{aligned}$$

where $W(b) = \text{diag}[\mu_i(b)\{1 - \mu_i(b)\}]$.

Finally, we obtain $b^{(t)}$, *i.e.*, the maximizer of $h(b)$. First, we rewrite $b^{(t)}$ as b_t; then we apply the Newton–Raphson method and obtain the iterative estimator of b_t as

$$b_t^{(k+1)} = b_t^{(k)} - [h''(b_t^{(k)})]^{-1} h'(b_t^{(k)}) = b_t^{(k)} + \left[Z^T W\left(b_t^{(k)}\right)Z + \frac{1}{\tau_G} I_L\right]^{-1} \left[Z^T\left\{Y - \mu\left(b_t^{(k)}\right)\right\} - \frac{b_t^{(k)}}{\tau_G}\right],$$

which depends on τ_G and γ, and we set $\tau_G = \widehat{\tau_G}^{(t)}$ and $\gamma = \gamma^{(t)}$. The maximizer, $b^{(t)}$, is obtained at each iteration until it converges, *i.e.*, until the difference $\left|b_t^{(k+1)} - b_t^{(k)}\right|$ falls below a prespecified threshold, *e.g.*, 10^{-7}. We denote the maximizer as $b_t^{(\infty)}$ and also set $b^{(t)} = b_t^{(\infty)}$.

Maximization step for obtaining $\hat{\gamma}^{(t+1)}$

To obtain $\hat{\gamma}^{(t+1)}$, we focus on the first term of $Q(\theta|\theta^{(t)})$; *i.e.*,

$$E\left\{\log f(Y|b;\theta)|Y;\theta^{(t)}\right\} = \sum_{i=1}^{n} E\left\{Y_i \log \mu_i + (1-Y_i)\log(1-\mu_i)|Y,\theta^{(t)}\right\} \equiv d(\gamma).$$

We rewrite μ_i as $\mu_i(\gamma)$ here to emphasize that it is a function of γ; *i.e.*, $\mu_i(\gamma) = \exp(X_i\gamma + Z_i b)/(1+\exp(X_i\gamma + Z_i b))$. We have that $\dot{\mu}_i(\gamma) \equiv \partial\mu_i(\gamma)/\partial\gamma = X_i^T(\exp(X_i\gamma + Z_i b)/\{1+\exp(X_i\gamma + Z_i b)\}^2) = X_i^T \mu_i(\gamma)\{1-\mu_i(\gamma)\}$. Then

$$\begin{aligned}
d'(\gamma) = \frac{\partial d(\gamma)}{\partial\gamma} &= \frac{\partial \sum_{i=1}^{n} E\left\{Y_i \log\mu_i + (1-Y_i)\log(1-\mu_i)\middle|Y,\theta^{(t)}\right\}}{\partial\gamma} \\
&= \sum_{i=1}^{n} E\left\{X_i^T Y_i \frac{\dot{\mu}_i(\gamma)}{\mu_i(\gamma)} + X_i^T(1-Y_i)\frac{-\dot{\mu}_i(\gamma)}{1-\mu_i(\gamma)}\right\} \\
&= \sum_{i=1}^{n} E\left\{X_i^T Y_i\{1-\mu_i(\gamma)\} - X_i^T(1-Y_i)\mu_i(\gamma)\right\} \\
&= \sum_{i=1}^{n} X_i^T(Y_i - \mu_i(\gamma)) \\
&= X^T(Y - \mu(\gamma)),
\end{aligned}$$

where $\mu = (\mu_1(\gamma), \mu_2(\gamma), \ldots, \mu_n(\gamma))^T = \mu(b)$, and

$$d''(\gamma) = \frac{\partial d'(\gamma)}{\partial\gamma^T} = \sum_{i=1}^{n} X_i^T(Y_i - \dot{\mu}_i(\gamma)) = -\sum_{i=1}^{n} \mu_i(\gamma)\{1-\mu_i(\gamma)\} X_i^T X_i = -X^T W(\gamma) X.$$

Recall that $W(\gamma) = \text{diag}\left\{\sum_{i=1}^{n}\mu_i(\gamma)\{1-\mu_i(\gamma)\}\right\} = \text{diag}\left\{\sum_{i=1}^{n}\mu_i(b)\{1-\mu_i(b)\}\right\} = W(b)$. Using the first and second derivatives of $d(\gamma)$, the estimator of $\gamma^{(t+1)}$, rewritten as γ_{t+1}, at the $(k+1)$th iteration, is given by

$$\begin{aligned}
\gamma_{t+1}^{(k+1)} &= \gamma_{t+1}^{(k)} - [d''(\gamma_{t+1}^{(k)})]^{-1} d'(\gamma_{t+1}^{(k)}) \\
&= \gamma_{t+1}^{(k)} + [X^T W\left(\gamma_{t+1}^{(k)}\right)X]^{-1} X^T\left(Y - \mu\left(\gamma_{t+1}^{(k)}\right)\right),
\end{aligned}$$

which depends on τ_G and b. We set $\tau_G = \widehat{\tau_G}^{(t)}$ and $b = b^{(t)}$. Then $\gamma^{(t+1)} = \gamma_{t+1}^{(\infty)}$.

Putting it all together, at iteration $t+1$ we have following estimators:

$\widehat{\tau_G}^{(t+1)} = -(1/r)[b^{(t)T}b^{(t)} + \text{trace}(\Sigma^{(t)})]$, where $b^{(t)} = b_t^{(\infty)}$ and $b_t^{(k+1)} = b_t^{(k)} + [Z^T W(b_t^{(k)})Z + (1/\tau_G)I_L]^{-1}[Z^T(Y - \mu(b_t^{(k)}) - b_t^{(k)}/\tau_G]$. $\gamma^{(t+1)} = \gamma_{t+1}^{(\infty)}$ and $\gamma_{t+1}^{(k+1)} = \gamma_{t+1}^{(k)} + [X^T \mu(\gamma_{t+1}^{(k)})X]^{-1} X^T(Y - \mu(\gamma_{t+1}^{(k)}))$.

Appendix C: Asymptotic Distributions of the Score Test Statistics

Recall that $T_{GE} = (1/2)\{(y_1^W - X\hat{\gamma})^T V_1^{-1} S_{GE} V_1^{-1}(y_1^W - X\hat{\gamma})\}\Big|_{\tau_G=\widehat{\tau_G},\ \tau_{GE}=0}$. Because $\hat{\gamma} = (X^T {V_0}^{-1} X)^{-1} X^T {V_0}^{-1} {Y_1}^W$, we have

$$\begin{aligned}
y_1^W - X\hat{\gamma} &= [I_n - X(X^T V_1^{-1} X)^{-1} X^T V_1^{-1}](y_1^W - X\gamma) \\
&= K_1(y_1^W - X\gamma),
\end{aligned}$$

where $K_1 = [I_n - X(X^T V_1^{-1} X)^{-1} X^T V_1^{-1}]$. Therefore, T_{GE} can be rewritten as

$$\begin{aligned} T_{GE} &= \frac{1}{2}\left\{(y_1^W - X\gamma)^T K_1^T V_1^{-1} S_{GE} V_1^{-1} K_1 (y_1^W - X\gamma)\right\} \\ &= \frac{1}{2}\left\{(y_1^W - X\gamma)^T V_1^{-1/2} V_1^{1/2} K_1^T V_1^{-1} S_{GE} V_1^{-1} K_1 V_1^{1/2} V_1^{-1/2} (y_1^W - X\gamma)\right\} \\ &= \frac{1}{2}\left\{\widetilde{y_1^W}^T A_1 \ \widetilde{y_1^W}\right\}, \end{aligned} \tag{C1}$$

where $\widetilde{y_1^W} = V_1^{-1/2}(y_1^W - X\gamma)$, and $A_1 = V_1^{1/2} K_1^T V_1^{-1} S_{GE} V_1^{-1} K_1 V_1^{1/2}$. In addition, the working vector y_1^W has mean $X\gamma$ and variance V_1 (Zhang and Lin 2003), and thus $\widetilde{y_1^W}$ has mean 0 and variance $I_{n \times n}$.

Let η_i^1, $i = 1, \dots, L$, denote the nonzero eigenvalues of matrix A_1 and let $\boldsymbol{\nu}_i^1$ denote the corresponding eigenvectors. Then, $T_{GE} = \Sigma_{i=1}^L \eta_i^1 (\boldsymbol{\nu}_i^{1T} \widetilde{y_1^W})^2 = \Sigma_{i=1}^L \eta_i^1 (Z_i)^2$, where $Z_i \dot{\sim} N(0,1)$. Therefore, T_{GE} can be approximated by a weighted sum of χ^2-distributions $\Sigma_{i=1}^L \widehat{\eta_i^1} \chi_{i(1)}^2$. By a similar derivation, the distribution of T_{joint} can be approximated by $\Sigma_{i=1}^L \widehat{\eta_i^0} \chi_{i(1)}^2$, where the η_i^0's are the nonzero eigenvalues of matrix $A_0 = V_0^{1/2} K_0^T V_0^{-1} (S_G + S_{GE}) V_0^{-1} K_0 V_0^{1/2}$, with $K_0 = [I_n - X(X^T V_0^{-1} X)^{-1} X^T V_0^{-1}]$.

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.171686/-/DC1

Assessing Gene-Environment Interactions for Common and Rare Variants with Binary Traits Using Gene-Trait Similarity Regression

Guolin Zhao, Rachel Marceau, Daowen Zhang, and Jung-Ying Tzeng

DOI: 10.1534/genetics.114.171686

File S1

Simulation code and data

Available for download as a .zip file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.171686/-/DC1